VISUAL FEATURES FOR MULTIMODAL SPEECH RECOGNITION

Petr Motlíček, Lukáš Burget and Jan Černocký
Faculty of Information Technology, Technical University Brno
Phone: +420-5-4114-1279, Fax: +420-5-4114-1270
E-mail: {motlicek,burget,cernocky}@fit.vutbr.cz

Abstract:

This paper demonstrates the use of visual parameters extracted from video for automatic recognition of phoneme strings. Encouraged by previous works utilizing "visually clean" data we investigate their efficiency in non-ideal conditions which are introduced by meeting audio-visual data employed in our experiments.

1 Introduction

Information from the speaker's mouth region has been shown to improve the accuracy and noise robustness of Automatic Speech Recognition (ASR) systems [1]. However, state-of-the-art systems use audio-visual speech features recorded under ideal lighting conditions. Those video recordings contain high-resolution video of subjects' frontal face, minimal changes in head positions and constant backgrounds. Insufficient research has been done in audio-visual ASR in real meeting conditions, where in addition to possibly noisy audio, the quality of visual channel is poor. In this paper, we describe preliminary experiments with audio-visual ASR system with visually challenging meeting data from real multi-party conversations.

2 Visual feature extraction

Prior to extraction of visual features, we need to process the original video stream data in order to detect and track faces (heads) of humans (objects) in meetings. The method employed is based on a skin color detection [2]. The visual input in our experiments is a video stream which is supposed to contain sequence of head poses of one human object. The video frame-rate $F_{vfr} = 25$Hz and the input resolution 70×70 pixel region is obtained for every video frame. Practically, each video frame contains the whole speaker's head including the hair and neck. A region-of-interest (ROI) is to be speaker's mouth, which should carry the most of variability caused by speech (it is of size 16×34 pixels) [2].

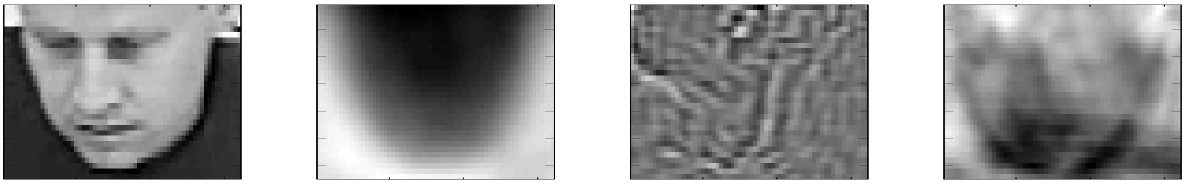

Figure 1: Application of PCA-LDA analysis: from left-to-right an original video frame (tracked head position), first PCA basis, first PCA-LDA basis, reconstructed video frame from first 10 PCA-LDA basis.

2.1 DCT coefficients

First, a two-dimensional, separable DCT is applied to the ROI and 16 lowest-order DCT coefficients are retained. The reasons for widespread use of DCT in feature extraction as well as in image compression, are the high compaction of the energy of the input signal onto a few DCT coefficients and the availability of a fast implementation of the transform, similar to the FFT. However, the DCT is not shift invariant, thus performance depends on a precise tracking of the ROI.

2.2 Methods for dimensional reduction and better class separability

The challenging problem in visual processing is the necessity to process highly dimensional data compared to audio signal. Therefore, some kind of transformation (i.e., DCT) needs to be applied. We were inspired by previous work of Potamianos et al. [4], which attempts to reduce dimensionality of input video frames and improves discriminability between classes while taking into account video training data. Widely used method to improve discriminability among classes is Linear Discriminant Analysis (LDA). However, its application directly on input video frames requires computation of large statistics (within-class and across-class covariance matrices) that would have to be estimated (in case of 32×64 video frames, the covariance matrices would be of size 2048×2048). Thus, first dimensionality reduction is provided by Principal Component Analysis (PCA). Execution of the following LDA is then computationally much less expensive.

3 Experimental setup

The audio-visual speech database collected at IDIAP [3] has been used. The parameters of audio speech signal are well-known Mel-filterbank log energies (23 banks, window size 20ms, audio frame-rate $F_{afr} = 100$Hz). Our main goal is the phoneme classification of audio-visual data based on Neural Nets (NN).

Derivation of visual features is the major part of our experiments. Each input video sequence is at the beginning processed by head tracking algorithm. Such visual data represents input to the following experiments:

A. 16 DCT coefficients from previously detected ROI are extracted (4 lowest DCTs in each dimension).

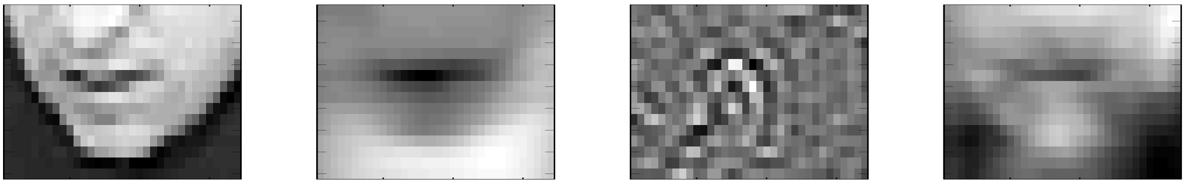

Figure 2: Application of PCA-LDA analysis: from left-to-right a mouth area (ROI) detected using correlation technique, first PCA basis, first PCA-LDA basis, reconstructed video frame from first 10 PCA-LDA basis.

B. We have performed several experiments with application of intra-frame LDA. A first-stage reduction is performed by PCA. Each video frame (32×64), containing lower half of tracked head position, is transformed into first 512 PCA basis. This data is used to estimate LDA statistics. Final transformation matrix is obtained by multiplying PCA and LDA matrices. Visual features used in recognizer are obtained by projection of input video frames onto first 45 PCA-LDA basis. Obviously, PCA and LDA statistics are estimated on training set. Fig. 1 shows original video frame, first PCA and PCA-LDA basis. For illustration, we also show a video frame reconstructed from first 10 PCA-LDA basis. Transformation obtained by LDA is not orthonormal, thus pseudo-inverse transformation is used for the reconstruction. It is not surprising that the first PCA basis returns global head pose estimated over all training data. On contrary to the PCA basis, PCA-LDA basis attempts to emphasize the discriminability between speech classes, which are in video frames mainly caused by large variation around objects' mouths.

C. We have also experimented with application of PCA-LDA transform directly onto ROI (mouth poses detected by a correlation-based approach). In this case, mouth regions (16×34) are used for estimation of PCA (256 basis) and PCA-LDA statistics. Finally, projection onto first 45 PCA-LDA basis is performed. The first PCA and PCA-LDA basis are shown in Fig. 2.

The acoustic and visual features are combined into a single vector which is then used in training and recognition processes. In order to cope with the different acoustic and visual frame-rate, visual parameters are upsampled up to 100Hz by a simple linear interpolation. Finally, acoustic and interpolated visual features are merged to build N-dimensional audio-visual feature vectors.

The evaluation of different audio-visual features was done on phoneme set that consists of 46 phonemes. The recognition system is a simple NN, as described in [2].

4 Experimental results

To evaluate our various audio-visual feature extraction algorithms, we observe the following results: a) the best frame-based phoneme accuracy on CV sets, b) a frame-based phoneme accuracy on forward passed test data. Experimental results are given in Tab. 1. During all experiments, the acoustic features were not touched. The vector size of visual features (as given in Tab. 1 together with acoustic features) did vary due to different kind of algorithms used to extract these parameters.

Feature extraction	CV [%]	FWD [%]	N
Audio only	28.9	31.0	23
DCT coefficients	28.33	31.33	23+16
PCA-LDA (head)	26.74	31.80	23+45
PCA-LDA (mouth)	27.59	**32.45**	23+45

Table 1: Experimental results - **frame-based phoneme accuracies**: CV - cross-validation set, FW - forward passed testing data, N - vector size of extracted audio-visual features.

5 Conclusions

This paper proposes a bimodal speech recognition scheme using visual parameters extracted from meeting recordings. The main goal is to combine such features with classical acoustic parameters in order to increase robustness of ASR. Experimental results related to the use of multimodal features are compared to the acoustic parameters (baseline). Many problems appear while processing video meeting data have not been properly solved yet. We need to improve the used head tracking algorithm as well as mouth detection method, which are still not very reliable, mainly due to real image conditions of visual data (low resolution of objects appearing in meetings, varying lighting conditions, etc.). We also expect that the temporal information from sequence of video frames, which has not been taken into account in our experiments yet, can play an important role.

6 Acknowledgments

This work was partially supported by EC project Multi-modal meeting manager (M4), No. IST-2001-34485, EC project Augmented Multi-party Interaction (AMI), No. 506811, and Grant Agency of Czech Republic under project No. 102/05/0278.

References

[1] G. Potamianos, C. Neti. "Automatic Speechreading of Impaired speech." *In Proceedings of Conf. Audio-Visual Speech Proc.*, pp.177-182, Aalborg, 2001.

[2] P. Motlíček, J. Černocký. "Multimodal Phoneme Recognition of Meeting Data." *Lecture Notes in Computer Science*, Springer-Verlag GmbH, Vol. 2004, Germany.

[3] D. Moore. "The IDIAP smart meeting room." *IDIAP Communication 02-07*, 2002.

[4] G. Potamianos, Ch. Neti. "Audio-Visual Speech Recognition in Challenging Environments." *In Proceedings of EUROSPEECH 2003*, pp. 1293-1296, Geneva, Switzerland, September 2003.